Analysis of multi-condition single-cell data with latent embedding multivariate regression

Constantin Ahlmann-Eltze[⋆,†] and Wolfgang Huber[⋆]

[⋆] *Genome Biology Unit, EMBL, Heidelberg, 69117, Germany.*
[†] *Collaboration for joint PhD degree between EMBL and Heidelberg University, Faculty of Biosciences*

6th of March 2023

Abstract

Multi-condition single-cell data reveal expression differences between corresponding cell subpopulations in different conditions. Current approaches divide cells into discrete groups or clusters and identify differentially expressed genes between corresponding groups. Here, we propose a method that operates without such grouping. *Latent embedding multivariate regression* (LEMUR) is based on a parametric mapping of latent space representations into each other and uses a design matrix to encode categorical and continuous covariates. We use the method to analyze a drug treatment experiment on brain tumor biopsies. We detect drug-induced gene expression responses affecting subsets of cells in a continuous latent space representation that does not require discrete categorization of the cells. Latent embedding multivariate regression is a versatile new approach for identifying differentially expressed genes from single-cell data of heterogeneous cell subpopulations or tissues under arbitrary experimental or study designs.

Contact: constantin.ahlmann@embl.de

Single-cell RNA-seq can be used to study the effect of experimental interventions or observational conditions on a heterogeneous set of cells, e.g., from tissue biopsies or organoids. Each unique combination of experimental or observational covariates is considered a *condition*. Typically, cells from the same sample (e.g., a biopsy) share the same condition but come from multiple cell types and states (e.g., position in a differentiation or cellular aging path, cell cycle, metabolism). There may be several samples (replicates) per condition. Compared to "bulk-sequencing", the novelty of single-cell RNA-seq is the ability to disentangle expression changes between corresponding cells (i.e., same cell type and state) under different conditions, from those between cell types or states.

This combination of explicitly known and latent covariates poses a challenge to regression or analysis of variance (ANOVA) methods. Variances observed in multi-condition single-cell data can be decomposed into four sources: the conditions, which are explicitly known or even set by the experimenter, cell type or state, which we consider a latent variable that is not explicitly given but can be inferred from the data with some degree of confidence and resolution, interactions between the two, and unexplained residual variability.

Currently, the prevalent approach is to convert the latent variable into discrete categories by unsupervised clustering and supervised classification. Thus, each cell is assigned to a cluster, and expression differences between such clusters across different conditions can be assessed using methods originally conceived for bulk RNA-seq data (Crowell et al., 2020). To ensure that the clustering is not confounded by the conditions (including technical covariates, also known as batches), methods for "harmonization" have been designed to integrate the data across conditions beforehand, including mutual nearest neighbors (Haghverdi et al., 2018), Harmony (Korsunsky et al., 2019), optimal transport (Schiebinger et al., 2019) and others. A recurrent challenge for harmonization methods is finding a balance

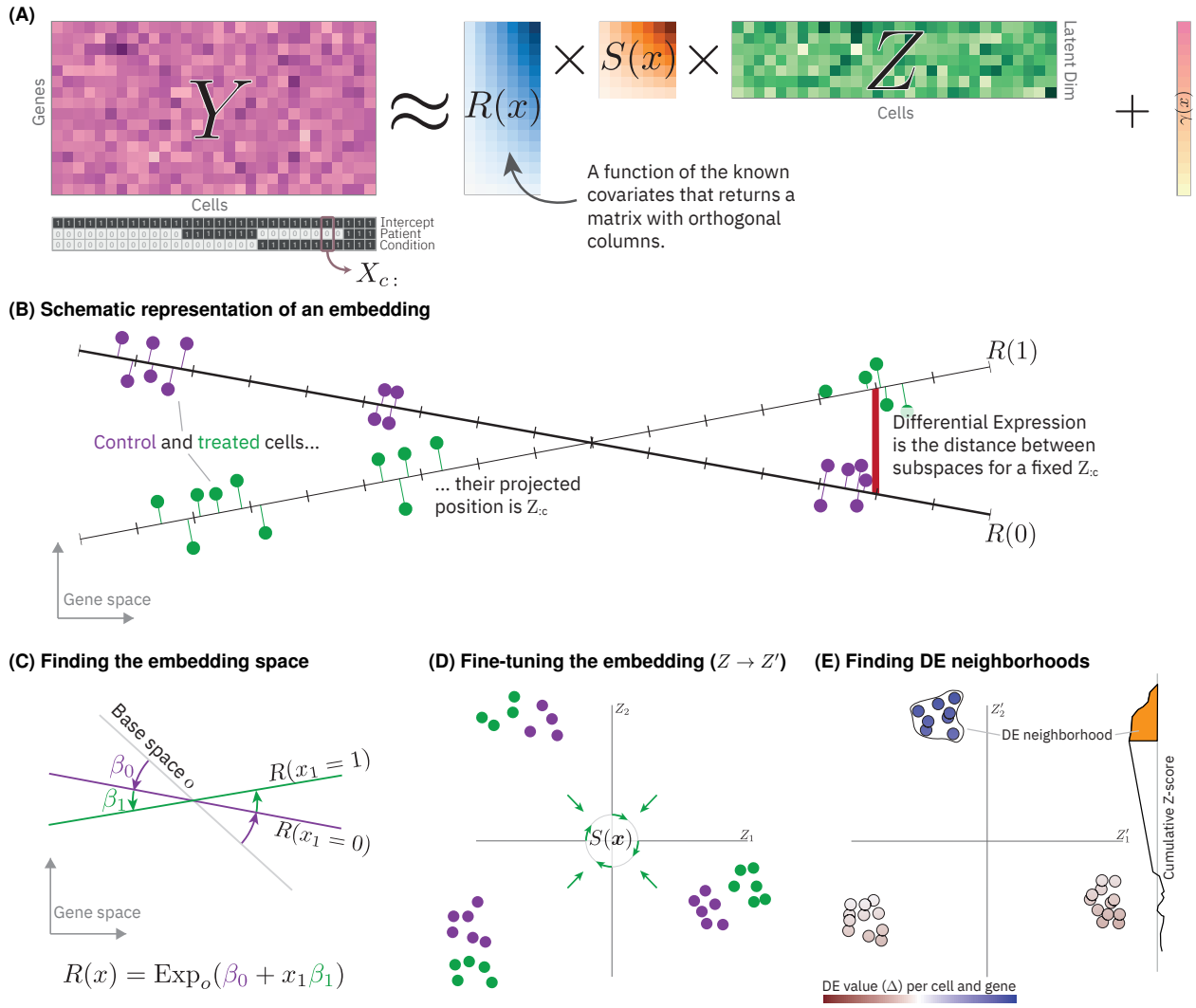


Figure 1 | Conceptual overview of latent embedding multivariate regression (LEMUR). (A) Graphical depiction of the matrix factorization at the core of LEMUR. (B) Comparison of cells from two conditions (green and purple), each with three subpopulations ("cell types"). LEMUR finds latent space representations, one for each condition (here, for practical reasons, the latent spaces are drawn as one-dimensional). (C) The latent spaces are produced by the function $R(x)$, which is parameterized by parameters β acting as high-dimensional rotations. (D) The function $S(x)$ does not affect the approximation of Y but changes the latent positions of the conditions relative to each other and can bring corresponding subpopulations closer. (E) Contrasting the predicted expression level from two conditions for each cell produces a differential expression (DE) value (Δ) for each gene and cell. We identify neighborhoods with consistently high (or low) DE values by calculating the cumulative z-score of the DE values along random directions. We call the group of cells with the maximum z-score a DE neighborhood.

between "correcting" unwanted variation and retaining wanted variation, i.e., biological signal of interest (Argelaguet et al., 2021).

The clustering-based approach has potential drawbacks. The most important one is that, while discrete cell types or states are a useful first-line abstraction, they may be an insufficient model of organismal biology for more sophisticated studies. Micro-environment, cell cycle, metabolic and paracrine differences can introduce gradual variability from cell to cell. A second drawback is that it is difficult to fully automate and tends to involve human intervention and judgement. Out of the box clustering algorithms can provide useful initial results, but reaching optimal clustering resolution is fiddly. Too small clusters mean insufficient power to detect changes, while too large clusters obscure

granular patterns. Rarer cell types, or those important to the biological question at hand may warrant more attention and higher resolution than others. Often, some degree of supervision is helpful, using reference expression profiles of previously annotated cell types (e.g., Aran et al. (2019)). All of these choices impact the differential expression analysis downstream in difficult to anticipate ways. In practice, this can generate a lot of back and forth. Thus, even if reporting results in terms of discrete cell types or states is a final objective, it would be more convenient if the manual human intervention and judgement step came more downstream in the workflow.

Here, we present a new statistical model for differential expression analysis (or ANOVA) of multi-condition single-cell data that combines the ideas of linear models and principal component analysis (PCA). The method, Latent Embedding MUltivariate Regression (LEMUR), is implemented in the `R` package *lemur*, which provides functions to assess the global effect of covariates on gene expression, to harmonize data from different conditions, to conduct cluster-free differential expression analysis, and to find cell neighborhoods that show consistent differential expression.

Results

LEMUR takes as input a data matrix Y of size $G \times C$, where G is the number of genes and C is the number of cells. The method assumes that appropriate preprocessing, including size factor normalization and variance stabilization, was performed (Ahlmann-Eltze and Huber, 2022). In addition, it expects specification of the design matrix X, of size $C \times K$ (Law et al., 2020). It produces several outputs:

- a low-dimensional representation of cells from all conditions,
- explicitly parameterized, bijective transformations that map the latent spaces into each other, and into a joint space,
- the predicted expression changes between any pair of conditions for each gene and cell, and hence the possibility to compute arbitrary contrasts, and
- neighborhoods of cells that show consistent differential expression.

We demonstrate the method on single-cell data from five glioblastomas that were cultured after surgical removal and treated using either vehicle control (DMSO) or panobinostat, an HDAC inhibitor (the data was originally collected and analyzed by Zhao et al. (2021)). We model these data using a paired control-treatment experimental design.

Regression of latent spaces

LEMUR is a matrix factorization algorithm and extends principal component analysis (PCA) (Fig. 1A). PCA (and similarly SVD) can be used to approximate a data matrix Y by a product of two simpler matrices

$$Y \approx RZ + \gamma_{\text{offset}}. \tag{1}$$

Here, R is a $G \times P$ matrix called *principal vectors* (or sometimes rotation or loadings matrix). The columns of R are orthonormal ($R^T R = I$). The $P \times C$ *embedding matrix* Z contains the P-dimensional coordinates of each cell in the latent space. If $P < \min(G, C)$, PCA reduces the dimension of the data. γ_{offset} is a vector with G rows and centers the observations[1].

LEMUR combines these ideas with regression analysis in the presence of covariates for the cells, which are encoded in the design matrix X. Instead of R being fixed, we treat it as a function of the covariates,

$$R : \mathbb{R}^K \to \{A \in \mathbb{R}^{G \times P} \mid A^T A = I_P\} \tag{2}$$

where the function arguments are rows of the design matrix and the output is the set of orthonormal $G \times P$ matrices. The details of the parametrization are explained in the Methods. Thus our model is

$$Y_{:c} \approx R(X_{c:})\, Z_{:c} + \gamma(X_{c:}), \tag{3}$$

where we use the notation : to indicate extracting row or column vectors from a matrix (e.g., $Z_{:c}$ is a vector of length P that contains the latent space representation of cell c). We allow the offset γ to depend on the covariates, too.

$R(x)$ is the latent space for all cells in condition x, i.e., all cells whose corresponding row in the design matrix equals x. This is illustrated in

[1] We overload the sum operator (+) for a matrix and a conformable column vector to produce another matrix: $C_{ij} = A_{ij} + b_i$

Fig. 1B, where we show a $G = 2$ dimensional gene expression space and a $P = 1$ dimensional latent space. In applications, the gene expression space has thousands of dimensions and typical choices for the latent space are $10 < P < 100$. Since R is defined on all of $\mathbb{R}^K$, the model can interpolate or extrapolate conditions that were not even measured.

Informally, we think of the function R in analogy to link functions in generalized linear models, which map linear predictors to statistical distributions from which observations are drawn. In our model, R maps the linear predictor for a cell to a linear subspace of the full gene expression space, in which we believe this cell's gene expression should lie (Fig. 1C).

Model (3) addresses the variance decomposition challenge posed in the introduction: known sources of variation are encoded in the design matrix X and act through the function $R(X)$, the latent variation (cell types or states) takes place in the linear space spanned by $R(X)$ and is parameterized by each cell's coordinates in Z. Interactions between the two are represented by condition-dependent changes in $R(x)$ that can differ in different directions of the embedding space Z, and unexplained variability is absorbed in the residuals of the approximation (Fig. 1B).

Fine-tuning the embedding

An assumption of Model (3) is that corresponding cell subpopulations from different conditions can be matched just by aligning their respective latent spaces through a high-dimensional rotation. Sometimes, this is not flexible enough, e.g., if a treatment drastically affects some, but not all cell subpopulations, and thus the relative distances between subpopulations change. To enable modeling of such localized changes, we extend our model by a condition-dependent linear alignment matrix S:

$$Y_{:c} \approx R(X_{c:})\, S(X_{c:})\, Z'_{:c} + \gamma(X_{c:}). \qquad (4)$$

The $P \times P$ matrix $S(x)$ is invertible and we define $Z'_{:c} \coloneqq S^{-1}(X_{c:})Z_{:c}$. This ensures that S only influences which subpopulations are considered "corresponding" and does not affect the approximation of Y. We find S by providing sets of cells that should have similar $Z'_{:c}$ (details in Methods).

Differential expression analysis

Model (4) predicts gene expression given a value of the covariates x and a position in the embedding space z. We calculate the differential expression for each gene and cell by comparing the predictions for any contrast of interest (e.g., between two conditions $x^{(A)}$ and $x^{(B)}$) for all Z' (Fig 1B,E).

The resulting matrix of differential expression values Δ $(G \times C)$ has two uses: first, we can visualize the values for selected genes as a function of latent space (in practice, we use for this a convenient 2D embedding of it, such as UMAP, McInnes et al. (2018)) to see how the differential expression possibly changes across that space. Second, we can use Δ to guide the identification of differential expression neighborhoods, i.e., cell types or states that are commonly showing differential expression for a particular gene (Fig. 1E, details in Methods). For statistical inference, we then use the established pseudobulking approach (Crowell et al., 2020) on that neighborhood and account for the statistical double dipping by count-splitting (Neufeld et al., 2022).

Analysis of a drug perturbation in glioblastoma

The glioblastoma study by Zhao et al. (2021) reported single-cell RNA-seq data of glioblastoma biopsies from five patients, each in two conditions: control and panobinostat, a non-selective histone deacetylase (HDAC) inhibitor. Fig. 2A shows the paired experimental design. There are 47 900 cells, and we considered the 6 000 most variable genes. We use the term *sample* for cells from one patient under one condition, so there are ten samples, and the number of cells per sample varies between 1 100 (2%, light purple) and 14 500 (30%, light blue) (Suppl. Tab. S1).

A two-dimensional visualization of the distribution of the cells by applying UMAP to the size factor normalized and shifted logarithm transformed matrix Y showed patterns most distinctively associated with the known covariates patient ID and treatment condition. There was further variation presumably related to different cell types (Fig. 2B). We used LEMUR to absorb patient and treatment effects into R, using a $P = 15$ dimensional latent space and fixing $S(x) = I$. Fig. 2C shows a UMAP of the matrix Z of latent coordinates for each cell. As

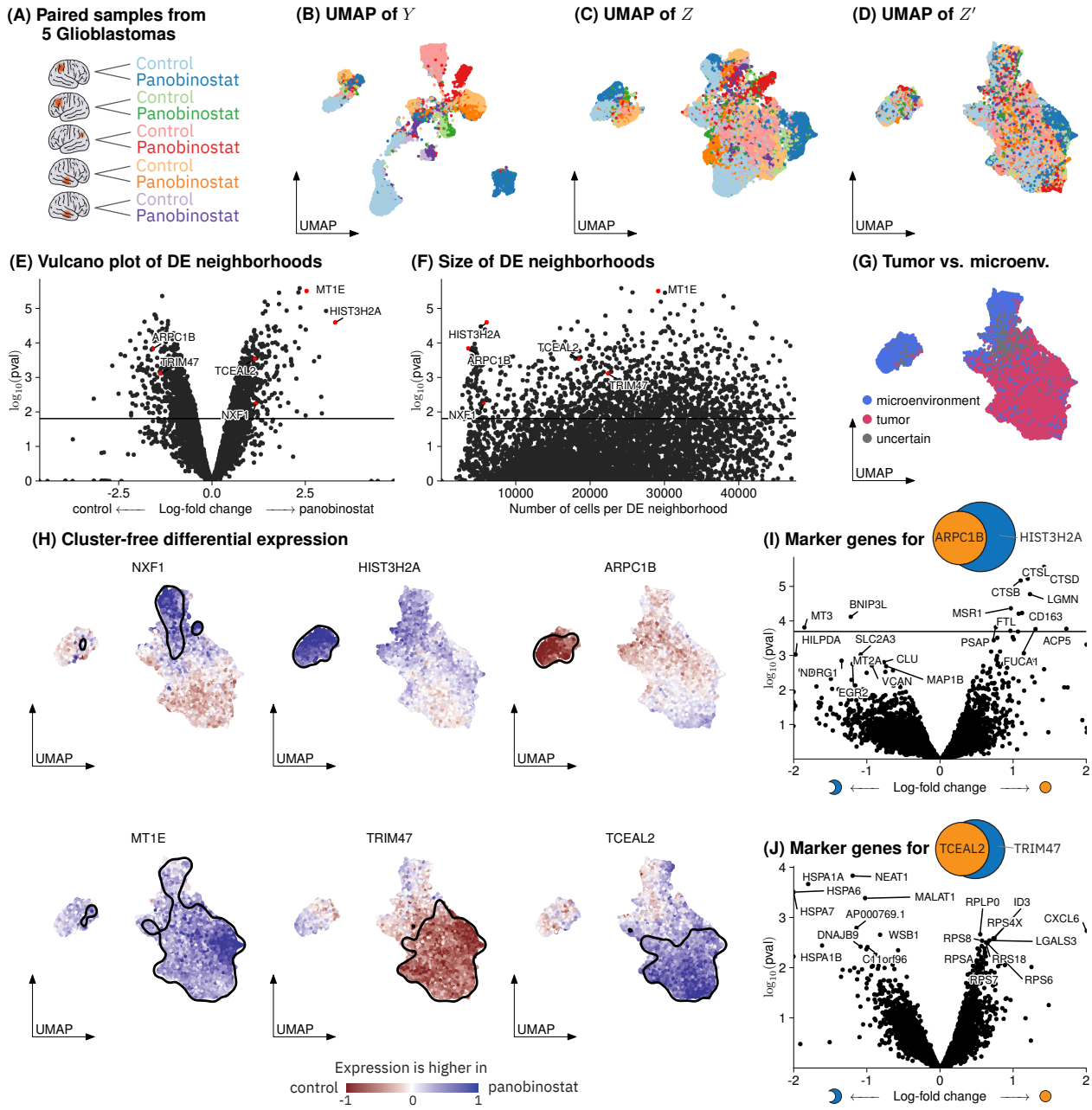


Figure 2 | Results from applying LEMUR to a glioblastoma dataset. (A) A schematic of the experimental design. (B-D) UMAP plots colored by condition on the input data: (B) Y, (C) the inferred latent position Z with $S(x) = I$, and (D) Z' after adjustment using maximum diversity clustering. (E) Volcano plots comparing panobinostat against control after forming pseudobulk samples per patient and condition for each differential expression neighborhood (one per gene). For the neighborhoods above the horizontal line, the FDR is 10%. (F) Same data as (E), but stratified by the differential expression neighborhood size. (G) UMAP of Z' colored by cell identity. Grey cells are of uncertain identity. (H) The differential expression inferred by LEMUR for six genes, with cells laid out by UMAP of Z'. The black boundary encircles 90% of the cells part of the differential expression neighborhood. Note that all distances and neighborhoods are evaluated in the 15-dimensional latent space and that the two-dimensional UMAP representations only serve for visualization. (I) Volcano plot for the pseudo-bulk comparison between the differential expression neighborhoods for ARPC1B and for HIST3H2A. (J) Same as (I) but for TCEAL2 and TRIM47.
The brain icon is by https://smart.servier.com, licensed under *CC-BY 3.0 Unported*, and was adapted for this figure.

a result, cells from different samples were more intermixed, and the visualization reflected more within-sample cellular heterogeneity. This picture became even clearer after we used S to encode an alignment between cell subpopulations across samples using Harmony's maximum diversity clustering (Fig. 2D). Here, a large tumor subpopulation (classified by Zhao et al. (2021) based on chromosome 7 amplification and chromosome 10 deletion) and two non-tumor subpopulations became apparent (Fig. 2G).

We predicted the expression change between panobinostat treatment and the control condition for all genes and cells. For each gene we identified exactly one differential expression neighborhood (details in Methods). More than 20% ($n = 1\,316$) of the differential expression neighborhoods showed significant up- or down-regulation in tests for differences between the pseudo-bulked counts (FDR = 10%, Fig. 2E). The large number of genes with a differential expression neighborhood is not surprising, as panobinostat is known for its potency and unspecific effects on gene expression (Atadja, 2009). For comparison, even the more unspecific approach of testing for differential expression across all cells identified a similar number of significant hits ($n = 1\,485$). The size of the differential expression neighborhoods varied from only a few hundred cells to encompassing almost all cells (Fig. 2F).

LEMUR identified biologically meaningful differential expression neighborhoods that matched evident subpopulations. We highlight six genes with significant expression changes to demonstrate the variety of differential expression patterns (Fig. 2H); in Suppl. Fig. S1, we show the underlying expression values. The differential expression patterns mostly corresponded to the cell subpopulations evident in the UMAP plot: e.g., upregulation of *NXF1* (nuclear RNA export factor 1) was predominantly in the non-tumor cells that express oligodendrocyte markers (Suppl. Fig. S2A). Similarly, the up-regulation of *HIST3H2A* and the down-regulation of *ARPC1B* was predominant in those non-tumor cells that express macrophage markers (Suppl. Fig. S2B). *ARPC1B* has been linked to the infiltration of tumor-associated macrophages in glioblastoma (Liu et al., 2022). Panobinostat treatment reduced the expression of *TRIM47*, which has been linked to inhibition of glioma proliferation (Chen et al., 2020), specifically in tumor cells.

LEMUR also identified biologically meaningful differential expression neighborhoods that did not correspond to an obvious subpopulation. The differential expression neighborhood of *ARPC1B* was almost completely contained in, but smaller than that of *HIST3H2A*; to find out if the difference between them was biologically meaningful, we looked at genes that distinguished the cells from the two sets in the control condition (Fig. 2I). The cells that were in both differential expression neighborhoods expressed many genes linked to tumor-associated macrophages: *CTSB*, *CTSD*, *CTLS* are peptidases linked to angiogenesis and tumor invasion (Olson and Joyce, 2015), *MSR1* is a macrophage marker, and *CD163* has been found upregulated in tumor-associated macrophages of the M2 (anti-inflammatory) phenotype (Komohara et al., 2008). In contrast, the cells that were only in the *HIST3H2A* differential expression neighborhood but not in that of *ARPC1B* showed increased expression of *MT3*, which has been associated with microglial cells (i.e., brain tissue resident macrophages) (Yoshiyama et al., 1998), *BNIP3L*, which is related to apoptosis (Imazu et al., 1999), and gene set enrichment analysis associated cellular response to hypoxia with the upregulated genes.

LEMUR identified two tumor subpopulations that consistently occurred across all five glioblastomas. When we contrasted the gene expression of the cells in *TCEAL2* differential expression neighborhood against that in cells from the *TRIM47* differential expression neighborhood, we found a clear pattern (Fig. 2J): The cells for the *TCEAL* differential expression neighborhood expressed more ribosomal genes, suggesting transcriptional activity, and chemokines linked to an immunosuppressive microenvironment (Wang et al., 2021). In contrast, the cells from the *TRIM47* differential expression neighborhood not in the *TCEAL2* neighborhood highly expressed many heat shock proteins, suggesting cellular stress. The patterns are not due to the overrepresentation of an individual patient in one of the neighborhoods (Suppl. Fig. S3). Note that although the changes are not statistically significant for individual genes (i.e., Benjamini-Hochberg adjusted p-values > 0.1), gene set enrichment analysis identified up-regulation of translation and downregulation of response to

unfolded proteins as significant.

Characterizing cells by their predicted expression change

To further explore the ability of LEMUR to identify cell subpopulations that respond similarly to a perturbation, we considered a dataset by McFarland et al. (2020), who measured the gene expression of 24 cancer cell lines before and after treatment with nutlin, an inhibitor of the interaction between MDM2 and p53. This drug is known to be only effective in *TP53* wild-type cells, which is the case for 7 of the cell lines. A UMAP visualization of the variance stabilized data showed two subpopulations for the *TP53* wild-type cells (colored) and no separation for the remaining cell lines (Fig. 3A). After adjusting for the nutlin perturbation with LEMUR, UMAP visualization separated only by cell line identity (Fig. 3B). In Fig. 3C, we visualize the predicted differential expression values for each gene and cell Δ for the top ten differentially expressed genes with UMAP. Consistent with the mechanism of action of nutlin, all the *TP53* mutated cell lines were merged into one cluster. This illustrates how the ability of LEMUR to predict the expression change between treated and untreated *for each cell* can facilitate the characterization of cells.

Discussion

We have introduced a method for the analysis of single-cell resolution expression data of heterogeneous tissues under multiple conditions with arbitrary experimental designs. LEMUR uses regression on latent spaces to enable cluster-free differential expression analysis. We have shown how it can harmonize data using linear transformations. We demonstrated its utility for finding differentially expressed genes and subsets of affected cells. Applied to the glioblastoma dataset by Zhao et al. (2021), LEMUR identified biologically relevant subpopulations and expression patterns.

Some aspects of the current implementation leave room for improvement: its last step, i.e., the inference of differential expression neighborhoods, can be sensitive to the choice of the dimension of the latent space. A second issue is the slow convergence of the method for designs with more conditions than covariates. Here, we usually stop the fitting after ten iterations, but more iterations or a more fundamental redesign of the optimization could improve the inference.

Overall, we believe that LEMUR is a valuable tool for first-line analysis of multi-condition single-cell data. Compared to approaches that require discretization into clusters or groups before differential expression analysis, representation of cell types and states in a continuous latent space may be a better fit to the underlying biology, which may enable discoveries that would otherwise be missed, or avoid false discoveries that stem from over-segmentation. Compared to deep-learning based latent space approaches, its interpretable, simple and easy-to-inspect model should facilitate follow-up investigation of its discoveries.

Availability

All datasets used in this manuscript are publicly available: the glioblastoma data is available on GEO at GSE148842, the cancer cell line data at figshare.

The *lemur* `R` package is available at `https://github.com/const-ae/lemur`.

Acknowledgments

We thank Dr. Simon Anders, Dr. Oliver Stegle and Dr. Judith Zaugg for valuable feedback and discussions. We thank Dr. Ronny Bergmann for his advice on optimization on manifolds.

Funding

This work has been supported by the EMBL International Ph.D. Programme, by the German Federal Ministry of Education and Research (CompLS project SIMONA under grant agreement no. 031L0263A), and the European Research Council (Synergy Grant DECODE under grant agreement no. 810296).

References

Ahlmann-Eltze, C. and Huber, W. (2020). glmGamPoi: fitting Gamma-Poisson generalized linear models on single cell count data. *Bioinformatics*, 36(24):5701–5702.

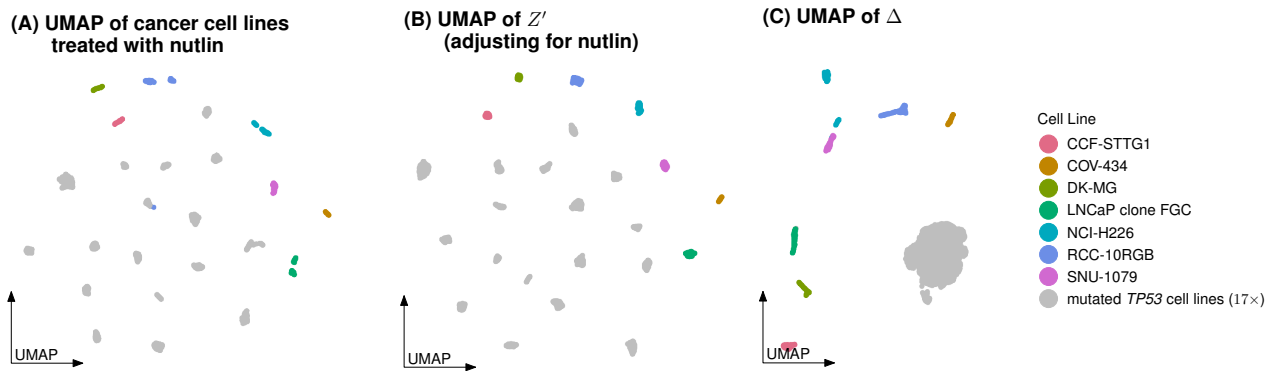


Figure 3 | Results from applying LEMUR to a cancer cell line dataset. (A) UMAP of 24 cancer cell lines treated with nutlin. The *TP53* wild-type cell lines are colored as they are receptive to nutlin treatment. (B) UMAP of the low-dimensional embedding from LEMUR adjusting for the nutlin treatment. (C) A UMAP on the differential expression matrix Δ for all cells and ten genes with the most variable differential expression values.

Ahlmann-Eltze, C. and Huber, W. (2022). Comparison of transformations for single-cell RNA-seq data. *bioRxiv*.

Aran, D., Looney, A., Liu, L., Wu, E., Fong, V., Hsu, A., Chak, S., Naikawadi, R., Wolters, P., Abate, A., Butte, A., and Bhattacharya, M. (2019). Reference-based analysis of lung single-cell sequencing reveals a transitional profibrotic macrophage. *Nature Immunology*, 20:163.

Argelaguet, R., Cuomo, A. S., Stegle, O., and Marioni, J. C. (2021). Computational principles and challenges in single-cell data integration. *Nature Biotechnology*, 39(10):1202–1215.

Atadja, P. (2009). Development of the pan-DAC inhibitor panobinostat (LBH589): Successes and challenges. *Cancer Letters*, 280(2):233–241.

Bendokat, T., Zimmermann, R., and Absil, P.-A. (2020). A Grassmann manifold handbook: Basic geometry and computational aspects. *arXiv preprint arXiv:2011.13699*.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, 57(1):289–300.

Chanzuckerberg Initiative (2023). CZ CELLxGENE Discover. `https://cellxgene.cziscience.com/`. Accessed: 2023-03-02.

Chen, L., Li, M., Li, Q., Xu, M., and Zhong, W. (2020). Knockdown of TRIM47 inhibits glioma cell proliferation, migration and invasion through the inactivation of Wnt/β-catenin pathway. *Molecular and Cellular Probes*, 53:101623.

Crowell, H. L., Soneson, C., Germain, P.-L., Calini, D., Collin, L., Raposo, C., Malhotra, D., and Robinson, M. D. (2020). Muscat detects subpopulation-specific state transitions from multi-sample multi-condition single-cell transcriptomics data. *Nature Communications*, 11(1):6077.

Haghverdi, L., Lun, A. T., Morgan, M. D., and Marioni, J. C. (2018). Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors. *Nature Biotechnology*, 36(5):421–427.

Imazu, T., Shimizu, S., Tagami, S., Matsushima, M., Nakamura, Y., Miki, T., Okuyama, A., and Tsujimoto, Y. (1999). Bcl-2/E1B 19 kDa-interacting protein 3-like protein (Bnip3L) interacts with bcl-2/Bcl-xL and induces apoptosis by altering mitochondrial membrane permeability. *Oncogene*, 18(32):4523–4529.

Kim, H. J., Adluru, N., Collins, M. D., Chung, M. K., Bendlin, B. B., Johnson, S. C., Davidson, R. J., and Singh, V. (2014). Multivariate general linear models (MGLM) on Riemannian manifolds with applications to statistical analysis of diffusion weighted images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2705–2712.

Komohara, Y., Ohnishi, K., Kuratsu, J., and Takeya, M. (2008). Possible involvement of the M2 anti-inflammatory macrophage phenotype in growth of human gliomas. *The Journal of Pathology: A Journal of the Pathological Society of Great Britain and Ireland*, 216(1):15–24.

Korsunsky, I., Millard, N., Fan, J., Slowikowski, K., Zhang, F., Wei, K., Baglaenko, Y., Brenner, M., Loh, P.-r., and Raychaudhuri, S. (2019). Fast, sensitive and accurate integration of single-cell data with Harmony. *Nature Methods*, 16(12):1289–1296.

Law, C. W., Zeglinski, K., Dong, X., Alhamdoosh, M., Smyth, G. K., and Ritchie, M. E. (2020). A guide to creating design matrices for gene expression experiments. *F1000Research*, 9:1444.

Liu, T., Zhu, C., Chen, X., Wu, J., Guan, G., Zou, C., Shen, S., Chen, L., Cheng, P., Cheng, W., et al. (2022). Dual role of ARPC1B in regulating the network between tumor-associated macrophages and tumor cells in glioblastoma. *Oncoimmunology*, 11(1):2031499.

McFarland, J. M., Paolella, B. R., Warren, A., Geiger-Schuller, K., Shibue, T., Rothberg, M., Kuksenko, O., Colgan, W. N., Jones, A., Chambers, E., et al. (2020). Multiplexed single-cell transcriptional response profiling to define cancer vulnerabilities and therapeutic mechanism of action. *Nature Communications*, 11(1):4296.

McInnes, L., Healy, J., and Melville, J. (2018). UMAP: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*.

Neufeld, A., Gao, L. L., Popp, J., Battle, A., and Witten, D. (2022). Inference after latent variable estimation for single-cell RNA sequencing data. *Biostatistics*.

Olson, O. C. and Joyce, J. A. (2015). Cysteine cathepsin proteases: regulators of cancer progression and therapeutic response. *Nature Reviews Cancer*, 15(12):712–729.

Schiebinger, G., Shu, J., Tabaka, M., Cleary, B., Subramanian, V., Solomon, A., Gould, J., Liu, S., Lin, S., Berube, P., et al. (2019). Optimal-transport analysis of single-cell gene expression identifies developmental trajectories in reprogramming. *Cell*, 176(4):928–943.

Wang, Z., Liu, Y., Mo, Y., Zhang, H., Dai, Z., Zhang, X., Ye, W., Cao, H., Liu, Z., and Cheng, Q. (2021). The CXCL family contributes to immunosuppressive microenvironment in gliomas and assists in gliomas chemotherapy. *Frontiers in Immunology*, 12:731751.

Wu, T., Hu, E., Xu, S., Chen, M., Guo, P., Dai, Z., Feng, T., Zhou, L., Tang, W., Zhan, L., et al. (2021). clusterProfiler 4.0: a universal enrichment tool for interpreting omics data. *The Innovation*, 2(3):100141.

Yoshiyama, Y., Sato, H., Seiki, M., Shinagawa, A., Takahashi, M., and Yamada, T. (1998). Expression of the membrane-type 3 matrix metalloproteinase (MT3-MMP) in human brain tissues. *Acta Neuropathologica*, 96:347–350.

Zhao, W., Dovas, A., Spinazzi, E. F., Levitin, H. M., Banu, M. A., Upadhyayula, P., Sudhakar, T., Marie, T., Otten, M. L., Sisti, M. B., et al. (2021). Deconvolution of cell type-specific drug responses in human tumor tissue with single-cell RNA-seq. *Genome Medicine*, 13(1):82.

Methods

The input data are a $G \times C$ matrix $Y \in \mathbb{R}^{G \times C}$ of gene expression measurements for G genes on C cells. The cells may come from multiple biological conditions and replicates (e.g., from different tissue specimens or an organoid under different treatments and/or developmental stages). This information is provided explicitly in K covariates and stored in the design matrix $X \in \mathbb{R}^{C \times K}$. Cells may also differ due to latent (i.e., not explicitly coded in X) factors, such as different cell types or cell states. A primary objective of the presented method is to identify these, to assign cells to them in a quantitative, probabilistic manner, and to learn how the latent factors "interact" with the explicitly coded factors, using a suitable definition of "interact".

Our method extends the PCA decomposition

$$Y = R\,Z + \gamma_{\text{offset}} + \varepsilon \tag{5}$$

where we approximate Y with a $P < \min(G, C)$ dimensional decomposition. The basis $R \in \mathbb{R}^{G \times P}$ has P orthonormal columns and the embedding $Z \in \mathbb{R}^{P \times C}$ contains the low-dimensional position for each cell. The offset $\gamma_{\text{offset}} \in \mathbb{R}^{G}$ accounts for the mean of each gene. We find R, Z, γ_{offset} by minimizing the squared residuals

$$\sum_{g=1}^{G} \sum_{c=1}^{C} \varepsilon_{gc}^2. \tag{6}$$

Intuitively, PCA finds a P dimensional subspace that minimizes the distance to the observed data Y; Z is the orthogonal projection of the data Y on the space spanned by R.

We incorporate the known covariates for each cell by fitting not just a single matrix R and a single offset vector γ_{offset}, but treating them as smooth functions of the covariates,

$$\begin{aligned} R &: \mathbb{R}^K \to \{A \in \mathbb{R}^{G \times P} | A^T A = I_P\} \\ \gamma &: \mathbb{R}^K \to \mathbb{R}^G, \end{aligned} \tag{7}$$

where the function arguments are rows of the design matrix, and the output of $R(x)$ is the set of orthonormal $G \times P$ matrices. Eqn. (5) then becomes

$$Y_{:c} = R(X_{c:})\, Z_{:c} + \gamma(X_{c:}) + \varepsilon_{:c}. \tag{8}$$

We also replace the offset vector with a function that returns a different offset for each condition. Eqn. (8) can be considered a *multi-condition* extension of PCA.

Intuitively, this multi-condition PCA finds a function that generates a P dimensional subspace for each condition that minimizes the distance to the observed data in that respective condition; Z is the orthogonal projection of the data on the corresponding subspace.

Before we explain how to parameterize the function R, we need to introduce some background on differential geometry. The set of orthonormal matrices is called a Stiefel manifold. Grassmann manifolds are closely related to Stiefel manifolds, except that matrices with the same span are considered equal (Bendokat et al., 2020). Accordingly, the elements of the Grassmann manifold $\text{Gr}(G, P)$ are P-dimensional subspaces in a G-dimensional ambient space, which we represent using matrices with orthonormal columns (i.e., elements of a Stiefel manifold). Working on a Grassmann manifold ensures that $R(x)$ is always a matrix with exactly orthonormal columns and that the interpolation between two subspaces is minimal.

We parameterize $R(x)$ as follows

$$R(x) = \text{Exp}_o\left(\sum_k x_k B_{::k}\right), \tag{9}$$

where the function argument x usually is a row from the design matrix and B are the coefficients, a 3-dimensional tensor of size $G \times P \times K$. The expression Exp_o is the exponential map on the Grassmann manifold. It takes a base point $o \in \text{Gr}(G, P)$ and a tangent vector $v \in T_o\,\text{Gr}(G, P)$ from the tangent space at point o, and returns a new point on the Grassmann manifold. The name exponential map derives from the fact that for some Riemannian manifolds the exponential map coincides with the matrix exponential; however, this is not the case for Grassmann manifolds. Here the exponential map for a base point o and a tangent vector A is

$$\begin{aligned} \text{Exp}_o^{(\text{Gr})}(A) = & o\,V\,\text{diag}(\cos(d))V^T \\ & + U\,\text{diag}(\sin(d))V^T, \end{aligned} \tag{10}$$

where $A = U\,\text{diag}(d)V^T$ comes from the singular value decomposition of the tangent vector.

Inside the exponential map (Eqn. (9)), we take linear combinations of the slices of B. Each slice of $B_{::k}$ is in the tangent space ($B_{::k} \in \mathcal{T}_o\text{Gr}(G, P)$) and we use the fact that any linear combination of elements from the tangent

Table 1 | Notation used in the manuscript.

Symbol	Meaning
Y'	Raw count data ($Y' \in \mathbb{Z}_{\geq 0}^{G \times C}$).
Y	Data after size factor normalization and variance-stabilizing transformation ($Y \in \mathbb{R}^{G \times C}$).
X	Design matrix ($X \in \mathbb{R}^{C \times K}$).
Z	Position of each cell in the low-dimensional embedding ($Z \in \mathbb{R}^{P \times C}$).
Δ	Predicted differential expression values between two conditions ($\Delta \in \mathbb{R}^{G \times C}$)
P	The number of embedding dimensions (akin to the number of dimensions in PCA). The index variable is p.
K	The number of covariates. The index variable is k.
G	The number of genes. The index variable is g.
C	The number of cells. The index variable is c.
$\gamma(x)$	Function that a vector of covariates (a row of X) and returns a vector with the offset for each gene $\gamma(x) \in \mathbb{R}^G$).
Γ	Linear coefficients of $\gamma(x)$ ($\Gamma \in \mathbb{R}^{G \times K}$).
$R(x)$	Function that takes a vector of covariates (a row of X) and returns a matrix with orthonormal columns ($R(x) \in \text{Stiefel}(G, P)$). The function generalizes the rotation matrix in PCA to multiple conditions.
B	3D-tensor of parameters that determine $R(x)$. Each slice $\beta_k = B_{::k}$ is an element of the tangent space $\mathcal{T}_o \,\text{Stiefel}(G, P)$.
o	The base-point ("zero") for the Grassmann exponential map (stored as $o \in \text{Stiefel}(G, P)$).
$S(x)$	Function that takes a vector of covariates and returns an invertible matrix ($S(x) \in \mathbb{R}^{P \times P}$).
$W^{(\text{rot})}$	3D-tensor of parameters for $S(x)$. Each slice $W^{(\text{rot})}_{::k} \in \mathcal{T}_I \,\text{Rotation}(P, P)$.
$W^{(\text{SPD})}$	3D-tensor of parameters for $S(x)$. Each slice $W^{(\text{SPD})}_{::k} \in \mathcal{T}_I \,\text{SPD}(P, P)$.

space remains in the tangent space (i.e., a tangent space is a vector space).

We parameterize the offset function $\gamma(x) = \sum_k \Gamma_{:k} x_k$, where $\Gamma \in \mathbb{R}^{G\times K}$. Accordingly, γ is just classical linear regression.

Fine-tuning the embedding

Multi-condition PCA (Eqn. (8)) only considers the subspaces spanned by each condition and does not consider the distribution of cells within that subspace. This makes it robust against overfitting, but the rigidity can also be limiting. We extend Model (8) with an extra term S, a non-distance preserving, linear isomorphism of $\mathbb{R}^P$, to (i) obtain additional flexibility and (ii) enable input of prior knowledge and user preferences in cell matching:

$$Y_{:c} = R(X_{c:})S(X_{c:})Z'_{:c} + \gamma(X_{c:}) + \varepsilon_{:c}. \quad (11)$$

Here, $Z'_{:c} := S^{-1}(X_{c:})Z_{:c}$. The extra term $S(x)$ distinguishes Eqn. (11), the LEMUR model, from its special case for $S \equiv I$, multi-condition PCA, Eqn. (8).

Next, we describe the selection of S, which is designed to enable the analyst to state preferences which cells from different conditions should be considered *similar*. We expect such a specification as a list of sets, each containing indices of cells to be considered similar across conditions. This can, for example, be derived from a set of matching cell type annotations, the set of mutual nearest neighbors, or Harmony's maximum diversity clustering. We denote the e-th set ($e = 1, \ldots, E \in \mathbb{N}$) as $\mathbb{E}_e$. This provision of preferences is optional; if it is lacking, we simply revert to $S \equiv I$, the identity, and to multi-condition PCA. If it is provided, S is obtained as a solution to the optimization problem

$$\underset{S\in\mathcal{S}(x)}{\arg\min} \sum_{e=1}^{E} \sum_{c\in\mathbb{E}_e} (M_e - S^{-1}(X_{c:})Z_{:c})^2, \quad (12)$$

where the optimization domain $\mathcal{S}(x)$ is described in the next paragraph, and $M_e = |\#\mathbb{E}_e|^{-1} \sum_{c\in\mathbb{E}_e} Z_{:c}$ is the mean latent space coordinate of the cells in similarity set e.

The optimization domain $\mathcal{S}(x)$, that is, the set of possible $S(x)$, is obtained from a multi-condition extension of the polar decomposition

$$S(x) = \mathrm{Exp}_I^{(\mathrm{SPD})}\left(\sum_k x_k W_k^{(\mathrm{SPD})}\right) \times \mathrm{Exp}_I^{(\mathrm{rot})}\left(\sum_k x_k W_k^{(\mathrm{rot})}\right), \quad (13)$$

where $\mathrm{Exp}^{(\mathrm{SPD})}$ is the exponential map of the $P\times P$ symmetric positive definite matrices (SPD) and $\mathrm{Exp}^{(\mathrm{rot})}$ is the exponential map of the $P\times P$ rotation matrices. Suppl. Fig. S4 gives a visual example how SPD and rotation matrices work. We implement a regularization that shrinks $S(X_{c:})$ towards the multi-condition PCA result by adding a ridge penalty for $W^{(\mathrm{SPD})}$ and $W^{(\mathrm{rot})}$ to Eqn. (12).

Implementation

The first step of fitting the LEMUR model is to choose the base space o, which serves as the reference or point of origin for the parameterization. We use the orthonormal matrix from computing PCA on all observations Y.

We fit multi-condition PCA, Eqn. (8) by repeatedly looping over the following steps:

1. Solve the linear regression for Γ, keeping $R(x)$ and Z fixed.
2. Optimize on the Grassmann manifold for the parameters B of the function $R(x)$, keeping Γ fixed.
3. Infer $Z_{:c}$ by projecting $Y_{:c}$ on the orthonormal basis $R(X_{:c})$.

In Step 2, we solve the manifold regression problem

$$\underset{B_{::k}\in\mathcal{T}_o\mathrm{Gr}(G,P)}{\arg\min} \left| Y_{:c} - \mathrm{Exp}_o\left(\sum_k B_{::k}X_{ck}\right) Z \right|^2 \quad (14)$$

by building on the work of Kim et al. (2014). They developed a generic algorithm to approximate the geodesic regression problem

$$\underset{B_{::k}\in\mathcal{T}_o\mathcal{M}}{\arg\min}\, d\left(\Omega_i, \mathrm{Exp}_o\left(\sum_k B_{::k}X_{ik}\right)\right), \quad (15)$$

where $\mathcal{M}$ is a generic manifold, $\Omega_i \in \mathcal{M}$ are data points on the manifolds, and

$$d(p,q) = \sqrt{\langle \mathrm{Log}(p,q), \mathrm{Log}(p,q)\rangle} \quad (16)$$

is the geodesic distance between two points on $\mathcal{M}$.

If the observations Ω_i are close to each other, the solution to Eqn. (15) is well approximated by the solution to a standard linear regression in the tangent space

$$\underset{B_{::k}\in\mathcal{T}_o\mathrm{Gr}(G,P)}{\arg\min} \left(\mathrm{Log}(o,\Omega_i) - \left(\textstyle\sum_k B_{::k}X_{ik}\right)\right)^2 \quad (17)$$

for a base point $o \in \mathcal{M}$ that is in the center of all Ω_i.

We cannot directly apply Kim et al. (2014)'s algorithm because our observations $Y_{:c}$ are not elements of the Grassmann manifold $\mathrm{Gr}(G,P)$. We resolve this problem as follows. We first construct a partition of $\{1,\ldots,C\}$, the set of all cells, into sets of cells that share the same condition: $\mathbb{D}_1 \cup \ldots \cup \mathbb{D}_D = \{1,\ldots,C\}$ and $\forall d \in 1,\ldots,D : \forall c_1, c_2 \in \mathbb{D}_d : X_{c_1:} = X_{c_2:}$. Then, for each group of cells under the same conditions (i.e., for each d) we find an orthonormal basis $U_d \in \mathrm{Gr}(G,P)$ using PCA on $Y_{:\mathbb{D}_d}$, the submatrix of Y for all cells from d. We then approximate a solution of Eqn. (14) by linear regression weighted by the number of observations per condition ($\#\mathbb{D}_d$) on the U_d projected into the tangent space of o.

Fine-tuning the embedding

We chose the parametrization in Eqn. (13) of $S(x)$ so we can easily express the inverse $S^{-1}(x)$. We selected the identity as the base point for the rotation and SPD exponential map because then both exponential maps reduce to the matrix exponential, and the inverse of the matrix exponential is just

$$\mathrm{Exp}(A)^{-1} = \mathrm{Exp}(-A). \quad (18)$$

Accordingly, the inverse $S^{-1}(x)$ is

$$\begin{aligned} S^{-1}(x) = &\mathrm{Exp}_I^{(\mathrm{rot})}\left(-\sum_k x_k W_k^{(\mathrm{rot})}\right) \times \\ &\mathrm{Exp}_I^{(\mathrm{SPD})}\left(-\sum_k x_k W_k^{(\mathrm{SPD})}\right). \end{aligned} \quad (19)$$

Using the expression $S^{-1}(x)$, we optimize the coefficients W^{SPD} and W^{rot} under the loss function Eqn. (12) analogous to the optimization of B applying the heuristic of Kim et al. (2014) iteratively until the algorithm converges.

Post-processing

After fitting the LEMUR model, we adjust the base space so that the rows of Z are sorted in descending order of their variance, i.e., we take our specific set of basis vectors and adjust them so that they correspond to the usual interpretation of principal components pointing in the direction of maximum variance. Specifically, we calculate a singular value decomposition of Z

$$Z = U\Sigma V^T. \quad (20)$$

We then set the base point to

$$\tilde{o} = o\,U, \quad (21)$$

adjust the coefficients of R to

$$\tilde{B}_{::k} = B_{::k}\,U \quad (22)$$

and set the low-dimensional embedding Z to

$$\tilde{Z} = \Sigma V^T. \quad (23)$$

Cluster-free differential expression

LEMUR learns a parametric model of the multi-condition single-cell data which we can use to predict expression changes between two conditions for each cell. If we use the inferred parameters for $\gamma(x)$, $R(x)$, and $S(x)$, we can write

$$f(x,z) = R(x)S(x)z + \gamma(x) \quad (24)$$

where f is a function that predicts the gene expression of a "virtual cell" at an arbitrary position z in the embedding space for any condition x.

Thus, the predicted differential expression for all genes in cell c between conditions A and B is

$$\Delta_{:c} = f(x^{(A)}, Z_{:c}) - f(x^{(B)}, Z_{:c}). \quad (25)$$

Differential expression neighborhoods

The differential expression matrix Δ guides the identification of neighborhoods that show consistent differential expression. These neighborhoods are gene-specific and we store them in a list $\mathbb{Q}$ of length G containing sets of the cell indices inside the neighborhoods.

To find the differential expression neighborhoods, we first sample many one-dimensional representations of the data Y. Specifically, we repeat the following many times: randomly sample

two cells from $\{1,\dots,C\}$ and calculate the connecting direction $v = Y_{:c_1} - Y_{:c_2}$. Then, project the data from all cells onto v, which results in a C-tuple $w \in \mathbb{R}^C$. We repeated this process often enough so there is a good chance that interesting differential expression patterns are apparent in one or more w's.

Next, we identify the best one-dimensional data representation for a gene g by choosing the w with the maximum absolute correlation to Δ_g. Intuitively, this selects a one-dimensional presentation of the data along which the differential expression varies.

We order the cells along w, calculate the cumulative z-score of $\Delta_{g:}$ in the new order, and use the neighborhood $\mathbb{Q}_g$ which has the maximum

$$z\text{-score} = \frac{\text{mean}(\Delta_{g\mathbb{Q}_g})}{\text{sd}(\Delta_{g\mathbb{Q}_g})/\sqrt{\#\mathbb{Q}_g}}. \tag{26}$$

Pseudobulk differential expression analysis

Pseudobulk samples aggregate the counts for all sample and subpopulation combinations. They effectively account for the fact that the experimental unit of replication in multi-condition single-cell data is the sample (and not the cells) (Crowell et al., 2020). The information which cell belongs to which sample creates a partition of $\{1,\dots,C\}$ into F sets of cells that we call $\mathbb{F}$. Here, we have to slightly modify the regular pseudobulk-formation procedure because we include a different set of cells in the pseudobulk for each gene.

Let $Y' \in \mathbb{Z}_{\geq 0}^{G\times C}$ be the count matrix based on which Y was constructed. Then we form the pseudobulk count matrix $V \in \mathbb{Z}^{G\times F}$ as

$$V_{gf} = \sum_{c\in\mathbb{F}_f\cap\mathbb{Q}_g} Y'_{gc}, \tag{27}$$

and calculate a gene-specific size factor

$$\text{sf}_{gf} = \sum_{c\in\mathbb{F}_f\cap\mathbb{Q}_g}\sum_{g'}^{G} Y'_{g'c}. \tag{28}$$

Relation with interaction models

The model in Eqn. (8) infers potential interactions between known covariates and the latent position of each cell. For example, a drug perturbation might affect the gene expression of cells early in a developmental trajectory more than in mature cells. Our model simultaneously identifies the latent position and the interacting drug effect. Yet, the way the interactions are modeled here differs from that in classical linear model interaction terms.

Conventional interactions are formed using a direct (Hadamard) product between two or more known covariates. For example, the effectiveness of trastuzumab on breast cancer cells depends on their HER2 status, i.e., the drug is more effective if the HER2 protein level is high. Accordingly, we could model cell viability as

$$\hat{y} = \beta_0 + \beta_1\, x_{\text{conc.}} + \beta_2\, x_{\text{HER2}} + \beta_3\, x_{\text{conc.}} \odot x_{\text{HER2}} \tag{29}$$

and call β_3 the interaction coefficient. Such a "classical" interaction model can be understood as an alternative specification of the function $R(x)$

$$R(x) = B\ (I_P \otimes x), \tag{30}$$

where B is a $G \times PK$ matrix, $\otimes$ is the Kronecker product,

$$X = \begin{pmatrix} | & | \\ 1 & x_{\text{conc.}} \\ | & | \end{pmatrix}, \tag{31}$$

$$Z = \begin{pmatrix} | & | \\ 1 & x_{\text{HER2}} \\ | & | \end{pmatrix}^T, \tag{32}$$

where the vertical bars indicate column-vectors of length C, and $P = 2$ and $K = 2$.

When we plug Eqn. (30) into $\hat{Y}_{:c} = R(X_{c:})Z_{c:}$, we can rewrite it as

$$\hat{Y} = \sum_{p,k} B_{:(k+pK)}\ (X_{:k} \odot Z_{p:}) \tag{33}$$

which emphasizes the relation to the classical interaction model.

Independent of the parametrization (Eqn. (9) or Eqn. (30)), $R(x)$ can be interpreted as spanning the space that best approximates the observations from condition x. The advantage of Eqn. (9) is that the constraints of the Grassmann manifold naturally map to this intuition. In contrast, the parametrization of Eqn. (30) does not enforce orthonormality between the columns of $R(x)$, it does not even enforce a common scale. This complicates interpreting and comparing the latent position $Z_{:c}$ for two cells.

Geometrically, the columns of B in Eqn. (30) that correspond to the intercept in X span a base space. All other columns in B are vectors that point out of that base space. In contrast, the $B_{::k} \in \mathcal{T}_o\mathrm{Gr}(G, P)$ in Eqn. (9) correspond to rotations of the base space. For small angles between the spaces of two conditions there is little difference between a rotation and the straight vector. Thus, one can interpret our multi-condition PCA model as approximating a conventional interaction model between observed and latent covariates.

Execution details

Glioblastoma analysis

To analyze the glioblastoma dataset (Zhao et al., 2021), we first split the counts into a test and a training set using *countsplit* (Neufeld et al., 2022) and set $\epsilon = 0.5$. Next, we accounted for the varying size factors per cell and variance stabilize the training counts using the `shifted_log_transformation` function from the *transformGamPoi* `R` package (Ahlmann-Eltze and Huber, 2022). We fit the LEMUR model using $P = 15$ and account for the patient ID and the treatment condition ($\sim$ `patient_id + treatment`). The fitting took 9 minutes for ten iterations on our cluster without parallelization.

We fine-tuned the alignment of the LEMUR model using Harmony's maximum diversity clustering, fitting one coefficient for each patient and treatment combination ($\sim$ `patient_id * treatment`). We set the regularization on $W^{(\mathrm{SPD})}$ to ∞, fitting only the rotation.

To identify the differential expression neighborhoods, we fit 100 random directions. We formed the pseudobulk of the test counts from countsplit, fit a Gamma-Poisson generalized linear model using *glmGamPoi* (Ahlmann-Eltze and Huber, 2020) accounting for patient ID and the treatment condition ($\sim$ `patient_id + treatment`) and tested the *panobinostat* vs. *control* condition (Ahlmann-Eltze and Huber, 2020). The resulting p-values were adjusted using the Benjamini-Hochberg method (Benjamini and Hochberg, 1995).

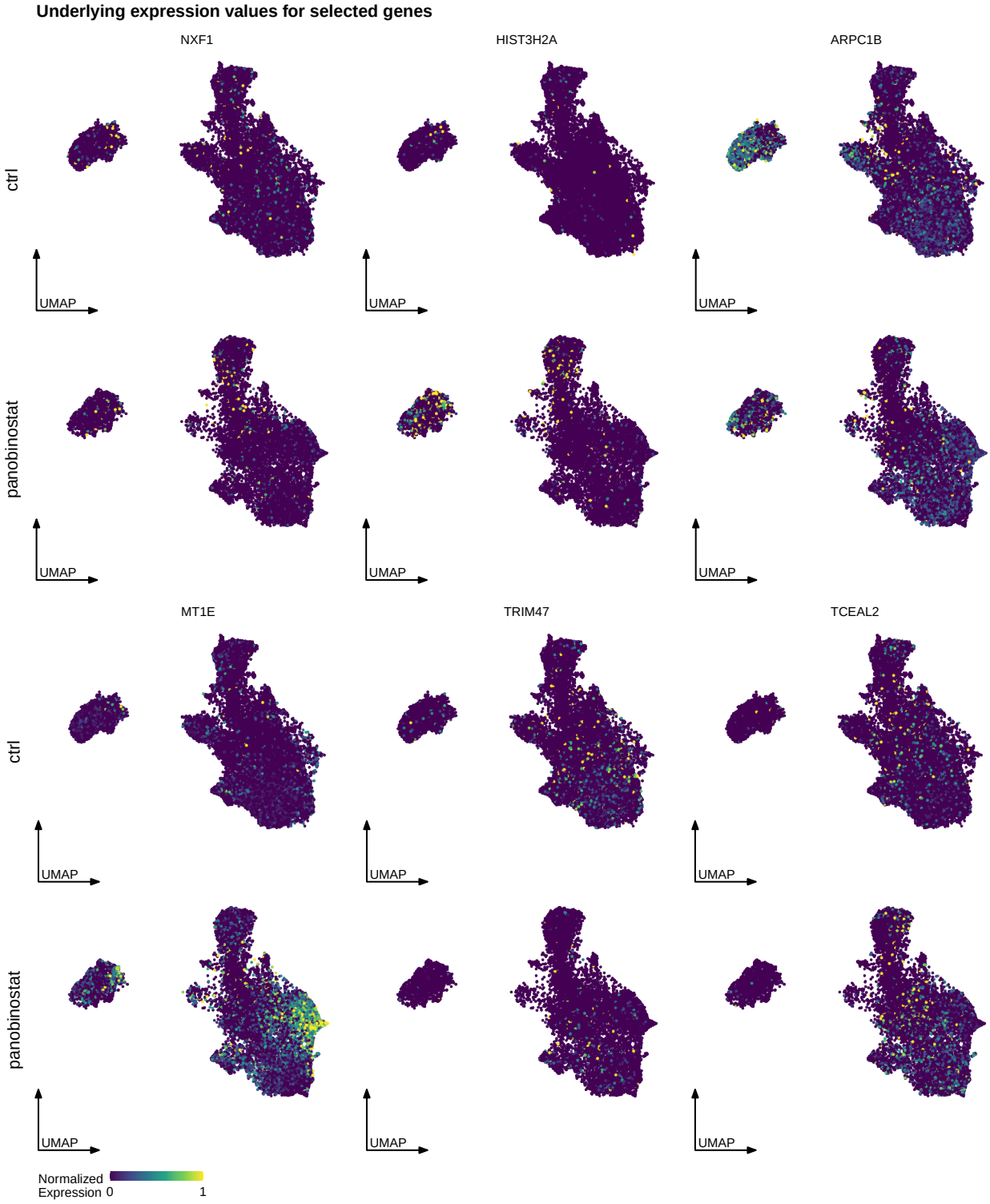
The color scale of the differential expression in Fig. 2H was normalized using $\Delta_{g:} / |\max\left(\mathrm{quantile}_{5\%,95\%}(\Delta_{g:})\right)|$, where the quantile function returns the 5% and 95% quantile of the differential expression vector. The shown boundary lines are the contour of the point density that comprises 90% of the points, calculated using *ggplot2*'s `stat_density` function.

The gene set enrichment analysis of the up and down-regulated genes from Fig. 2 was conducted using the `enrichGO` function from *clusterProfiler* (Wu et al., 2021). We used all 6 000 highly variable genes as background to test the over-representation of the 30 most significantly changed genes.

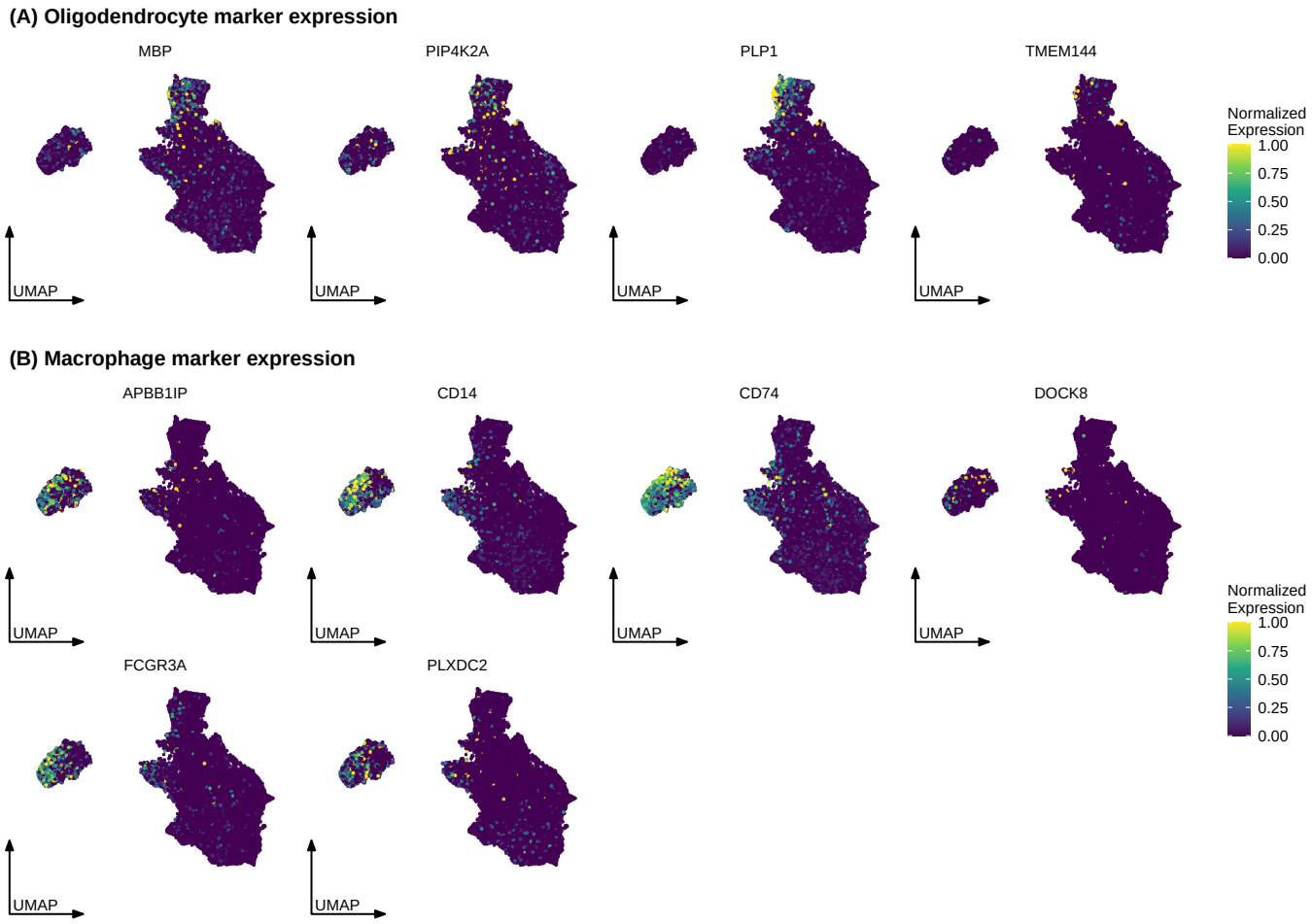
Cancer cell line analysis

We adjusted the cancer cell lines data from McFarland et al. (2020) for the varying size factors and variance stabilize the counts using *transformGamPoi*. We restricted our analysis to the $2,000$ most variable genes. We fitted a $P = 30$ dimensional model with LEMUR and accounted for the treatment condition. We fine-tuned the alignment using Harmony's maximum diversity clustering without regularization. We produced the UMAP of Δ matrix on the 10 most variable genes as measured by the row-wise variance of Δ.

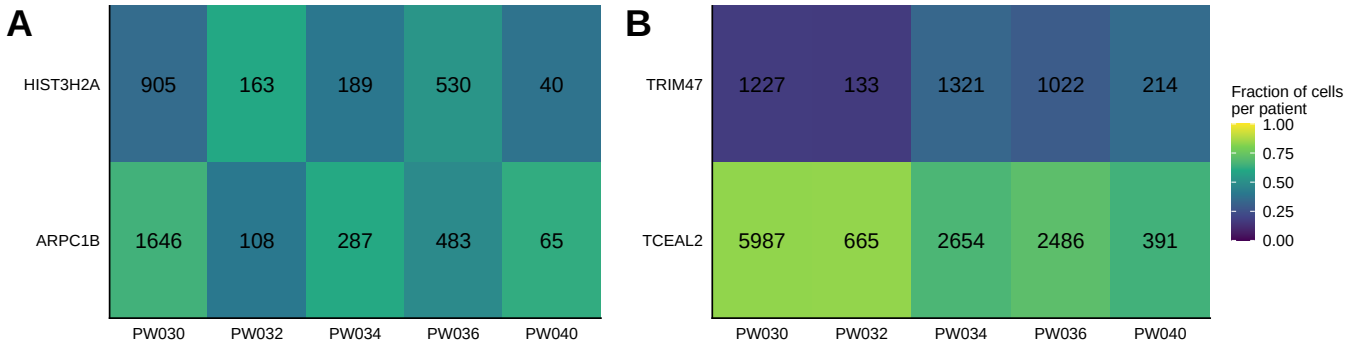
Supplementary Figures



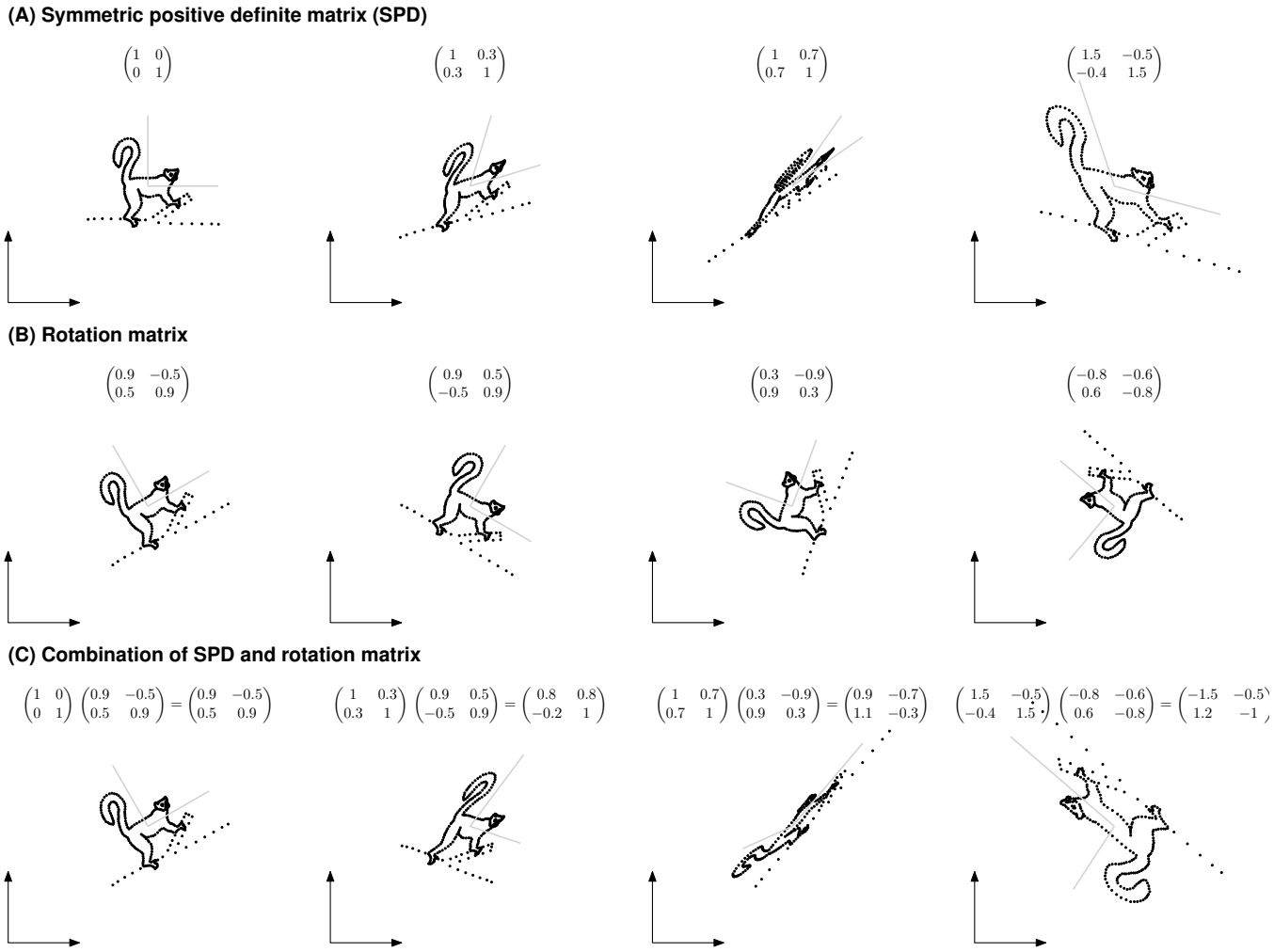
Suppl. Figure S1: Expression values for six selected genes in the control and panobinostat treated condition. Fig. 2H shows the cell-wise differential expression values inferred by LEMUR. The expression was brought to a common scale across genes by dividing the expression values by their 99% quantile.



Suppl. Figure S2: Expression patterns of selected marker genes on the UMAP of Z' for (A) oligodendrocytes and (B) macrophages. The gene sets are based on manual curation and the top marker according to the CZ CELLxGENE cell type annotations (Chanzuckerberg Initiative, 2023). The expression was brought to a common scale across genes by dividing the expression values by their 99% quantile.



Suppl. Figure S3: Number of cells per patient in the *ARPC1B* and the *HIST3H2A* differential expression neighborhood (A) and the *TCEAL2* and *TRIM47* differential expression neighborhood (B).



Suppl. Figure S4: Visualization of the effect of a symmetric positive definite (SPD) matrix (A), rotation matrix (B), and their combination (C) on a two-dimensional point cloud.

Suppl. Table S1: Overview of the glioblastoma patients.

Patient ID	Conditon	Age	Gender	Tumor location	# Cells
PW030	0.2 uM panobinostat	65	M	right parietal	7118
PW030	vehicle (DMSO)	65	M	right parietal	14478
PW032	0.2 uM panobinostat	61	M	left frontal	1401
PW032	vehicle (DMSO)	61	M	left frontal	1491
PW034	0.2 uM panobinostat	68	F	left parieto-occipital	2679
PW034	vehicle (DMSO)	68	F	left parieto-occipital	7782
PW036	0.2 uM panobinostat	56	M	right temporal	3096
PW036	vehicle (DMSO)	56	M	right temporal	6749
PW040	0.2 uM panobinostat	69	M	right temporal	1987
PW040	vehicle (DMSO)	69	M	right temporal	1119